Assimilating ocean color data using an iterative ensemble smoother: skill assessment for a suite of dynamical and error models

K.W. Smith and D.J. McGillicuddy, Jr.[1]

Manuscript submitted to Journal of Marine Systems

January 5, 2010

[1] Department of Applied Ocean Physics and Engineering, Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA. Tel: 508-289-2683 Fax: 508-457-2194 Email: dmcgillicuddy@whoi.edu (Corresponding Author).

Abstract

Inference of the sea surface chlorophyll field from incomplete satellite coverage is posed as a formal inverse problem using a Monte Carlo approach to Bayesian estimation. We introduce a new method, the strong constraint iterative ensemble smoother, for solving the general coupled physical-biological parameter estimation problem where model nonlinearities may be relevant. The forward model is posed in four ways: (1) advection-diffusion, (2) linear advection-diffusion-reaction, (3) nonlinear advection-diffusion-reaction, and (4) a nonlinear nutrient-phytoplankton model. Hindcast skill is demonstrated through analysis of the fit to independent data in a series of experiments utilizing MODIS chlorophyll imagery from the Middle Atlantic Bight during summer of 2006. The data assimilative model demonstrates skill over a range of presumed observational error. Both the purely physical model (advection-diffusion only) and the coupled physical-biological models exhibit skill fitting unassimilated data. The skill of the coupled physical-biological models is greater than the skill of the advection-diffusion model, owing at least in part to greater degrees of freedom in those inversions.

Key Words: Data assimilation, inverse models, Monte Carlo methods, physical-biological interactions, satellite ocean color

Regional terms: USA, Middle Atlantic Bight

1. Introduction

Common methods for compositing and interpolating satellite imagery typically rely on regression and smoothing of individual pixels, inherently ignoring the effect of advection. With improvements in shelf-scale observing systems and expanding areas of coverage by operational models, we are faced with the opportunity to improve sea surface chlorophyll (SSC) estimates. An analogous situation exists with respect to biological models. Although the dynamics of plankton ecosystems remain an active topic of research, direct contact between models and observations via biological data assimilation (Fennel et al., 2001; Hofmann and Friedrichs, 2002) is leading to demonstrable improvements in skill (Lynch et al., 2009). Herein we pose the SSC compositing problem as dynamic interpolation, formally inverting a model to fill in the gaps in the data.

In the data assimilation problems characteristic of today's ocean (spatially explicit models with millions of state variables assimilating hundreds of sparsely distributed data points) some type of Bayesian reasoning must be brought to bear to obtain a well-posed inverse problem. The prior information may enter as gradient or other penalty in a cost function or be explicitly stated as prior distributions on the parameters being estimated. A potential drawback to any Bayesian approach to data assimilation is that the analyst will bias the results with the specification of the prior error distributions. We seek to demonstrate robustness of an estimation procedure with respect to specification of the prior error distributions for several different models, using an example set of eleven sequential satellite images from the Middle Atlantic Bight during summer 2006.

Satellite sensed ocean color data has been assimilated by various methods. Ishizaka (1990) used a simple insertion-based methodology with Coastal Zone Color Scanner (CZCS) data. Natvik and Evensen (2003a; 2003b) assimilated Sea-viewing wide Field-of-view Sensor

(SeaWiFS) data into a three-dimensional plankton ecosystem model using a Ensemble Kalman Filter (EnKF). More recently Gregg (2008) assimilated SeaWiFS data into a global biogeochemical model using the Conditional Relaxation Analysis Method (CRAM) in a sequential manner. In all these applications the inverse problem is formulated in a weak constraint manner. Examples of ocean color assimilation using strong constraint formalism are comparatively few in number, generally making use of the adjoint method (e.g. Friedrichs (2002)). Seldom have such variational methods been applied in spatially explicit models (e.g. Garcia-Gorriz et al. (2003); Zhao and Lu (2008); Fan and Lv (2009)).

Monte Carlo ensemble methods offer an alternative approach, which can be formulated either in terms of weak constraint (Evensen, 2006; van Leeuwen and Evensen, 1996) or strong constraint (Smith et al., 2009). The ensemble smoother (EnS) holds two practical advantages over variational methods described above. Firstly, the implementation is vastly simpler because it does not require computation of the tangent linear model, which can be complicated for biological models. This allows for easy porting between applications to different biological models. Secondly, the method provides a Monte Carlo sample of the posterior error distribution without the need for computing the Hessian matrix. Posterior statistics are an important part of any estimation procedure, providing a context for assessing confidence in the conclusions. The ensemble smoother derivation relies on an assumption that the log likelihood is approximately quadratic (or, equivalently, that the model responds approximately linearly to the parameters at the observation points). This assumption can fail to hold depending on nonlinearities in the model, the oceanographic phenomenology present during the time period the data were collected, and the assumed observational error.

Herein we introduce a variation on the strong constraint Ensemble Smoother, the Iterative

Ensemble Smoother (ItEnS), for estimation problems in which the log likelihood is potentially strongly non-quadratic. The method utilizes a Monte Carlo approximation of the sensitivity matrix to provide the gradients for an iterative descent. Like the EnS, the iterative ensemble smoother does not require a tangent linear model. We apply this methodology to assimilating satellite-based ocean color data into four different dynamical models of varying complexity, and assess the performance of the algorithm and its dependencies on the underlying models and prescribed error statistics.

2. Methods

2.1 Forward Models

We investigate four coupled physical-biological models. The first model we consider is a simple advection-diffusion (AD) model with no active biological interactions:

$$\frac{dc}{dt} + v \cdot \nabla c - \nabla D \cdot \nabla c = 0 \qquad (1)$$

where c is the chlorophyll concentration, v is the velocity field and D the diffusion field. The circulation estimate (v) is prescribed from a hindcast of the region described in He and Chen (submitted) (Figure 1). The velocity is a monthly average and the mesh resolution is approximately 8.9 km. A uniform horizontal diffusion coefficient (D) of 25 m^2 s^{-1} is used throughout. The model represents the vertical average over the top 20 meters of the water column.

The second biological model is a simple advection-diffusion-source (ADS) equation,

$$\frac{\partial c}{\partial t} + v \cdot \nabla c - \nabla D \cdot \nabla c = S(x, y) \qquad (2)$$

where S is a spatially variable source-sink term. An imposition of positivity on c causes the model to be nonlinear.

Our third model is an advection-diffusion-reaction (ADR) model with first order density dependence,

$$\frac{\partial c}{\partial t} + v \cdot \nabla c - \nabla D \cdot \nabla c = R(x, y)c \qquad (3)$$

where R is a spatially variable growth/loss rate. This is the simplest nonlinear model for a single biological state variable.

The last model we consider is a nutrient-phytoplankton (NP) model with Lotka-Voltera interaction and constant mortality rate for the phytoplankton,

$$\frac{\partial n}{\partial t} + v \cdot \nabla n - \nabla \cdot D \nabla n = \upsilon c - \gamma n c \qquad (4)$$

$$\frac{\partial c}{\partial t} + v \cdot \nabla c - \nabla \cdot D \nabla c = \gamma n c - \upsilon c \qquad (5)$$

where n is the nutrient concentration and c is the phytoplankton concentration. Parameters γ and υ represent the nutrient uptake rate and phytoplankton mortality rate, respectively. For the NP model the chlorophyll observations are assumed to be linear measurements of the phytoplankton field. This measurement model could be improved by explicitly accounting for variations in chlorophyll per unit biomass that can occur in phytoplankton due to photoadaptation (e.g. Cullen, (1982)). However, that refinement is left for future work.

All of these models offer simple description of the satellite-based chlorophyll observations, differing in explicit biological assumptions. For the ADS model, the free

parameters are initial conditions for *c* and the source/sink term *S(x,y)*. Likewise for the ADR model, the unknowns are initial conditions for *c* and the growth/mortality rate *R(x,y)*. For the NP model, the free parameters are the initial conditions for the two state variables *n* and *c*, as well as values of the parameters γ and υ. The abiotic AD model has only the initial condition of *c* for free parameters. The forward models are solved with an implicit time stepping finite element method as described in Smith et al. (2009).

2.2 Bayesian parameter estimation

We formulate the data assimilation problem using Bayesian formalism to estimate the parameters of a dynamical model given a set of observations. Let θ denote the unknown model parameters: for the AD model $\theta = \{c(t=0)\}$, for the ADS model $\theta = \{c(t=0), S\}$, for the ADR model $\theta = \{c(t=0), R\}$, and for the NP model $\theta = \{n(t=0), p(t=0), \gamma, \upsilon\}$. Let $f(\theta)$ denote the prior distribution for the parameters, and ψ_θ the dynamical model solution given parameter choice θ. The data, d, are an imperfect observation of the true state of the system, $d = H\psi_{true} + \xi$ where ξ is the observational error and *H* is the measurement operator for the observations. Bayes theorem allows us to compute the posterior likelihood over θ,

$$f(\theta \mid d) = \frac{f(d \mid \theta) f(\theta)}{\int f(d \mid \theta) f(\theta) d\theta} \propto f(d \mid \theta) f(\theta) = f(d \mid \psi_\theta) f(\theta) \qquad (6)$$

We seek the maximum likelihood estimate of θ over this posterior distribution. Assuming $f(\theta)$ is Gaussian, let μ and P denote the prior mean and covariance of θ. If the observations are unbiased and perturbed by an additive Gaussian error distribution with covariance *W* and zero mean, then

$$f(\theta \mid d) = \frac{1}{\sqrt{(2\pi)^{N_m + N_d} |W||P|}} \exp\left(-\frac{1}{2}(H\psi_\theta - d)^T W^{-1}(H\psi_\theta - d) \right) \exp\left(-\frac{1}{2}(\theta - \mu)^T P^{-1}(\theta - \mu) \right) \quad (7)$$

where N_m is the dimension of the model and N_d is the dimension of the data. The analogy to strong constraint data assimilation methods is illuminated by defining a cost function proportional to the log of the conditional likelihood function,

$$J(\theta) \propto -2\log(f(\theta \mid d)) = (H\psi_\theta - d)^T W^{-1}(H\psi_\theta - d) + (\theta - \mu)^T P^{-1}(\theta - \mu) \quad (8)$$

By monotonicity of the log, the value of θ minimizing the cost is also the maximum likelihood estimate.

Bayesian methodology requires the specification of prior distributions for unknown parameters, $f(\theta)$, and observations, $f(d \mid \psi_\theta)$. In general, the prior distributions over the parameters and observations are specified by analytic functions with a handful of scalar parameters. Herein we refer to these parameters as "hyper-parameters," and their values for models describing geophysical systems are often not known with great confidence.

2.3 Observational error covariance

Above we asserted that the observations contain additive Gaussian errors with mean zero and covariance *W*. Generally *W* is assumed to be a constant diagonal matrix (measurement errors are not correlated and have the same expected error). For the satellite data used herein, we employ a block diagonal covariance

$$W_{ij} = \sigma_{obs}^2 \exp\left(-\frac{|x_i - x_j|}{l_{obs}} \right) \delta(t_i - t_j) \quad (9)$$

defining the covariance between observation *i* and *j*, where t_i and t_j are the times of the observations and x_i and x_j are the positions of the observations. The delta function, $\delta(t)$, is one at the origin and zero elsewhere. This form is based on the assumption that while the errors in separate images are uncorrelated, data within an image is contaminated with a spatially correlated signal. The decorrelation length scale of the observations l_{obs} was estimated directly from the satellite-based chlorophyll data (Table 2), and a range of values for the observational error σ_{obs} is investigated (see section 2.7 below).

2.4 Prior error distributions

We assume a Gaussian error distribution for the initial conditions in all of the models. The distribution is truncated to enforce positive definiteness in the initial conditions. The prior model distribution at later times is estimated through the solution of the forward models (Equations 1-5). The distributions of *S, R, n(t=0),* γ, and υ are also assumed to be Gaussian and independent of the initial conditions.

The covariance for the initial conditions varies spatially in proportion to the climatological mean field,

$$C_{ij}^{0} = \sigma_0^2 \mathrm{g(x_i)g(x_j)exp}\left(-\frac{|x_i - x_j|}{l_m}\right) \qquad (10)$$

where $g(x)$ is mean initial condition for chlorophyll provided by the MODIS August climatology. The nondimensional hyper-parameter σ_0 is the standard deviation of the chlorophyll data itself scaled by the climatological mean value. The length scale for model

errors l_m was also estimated from the observations by computing the decorrelation length scale of normalized chlorophyll anomalies $\frac{c-\bar{c}}{\bar{c}}$ (Table 2).

The covariance for the reaction term in the ADS and ADR models are:

$$C_{ij}^{S} = \sigma_S^2 \exp\left(-\frac{|x_i - x_j|}{l_m}\right) \text{and } C_{ij}^{R} = \sigma_R^2 \exp\left(-\frac{|x_i - x_j|}{l_m}\right) \quad (11)$$

respectively. The means for both *S* and *R* are zero, and variances were estimated from satellite imagery under the assumption of no flow (Table 2):

$$\sigma_S^2 = \sqrt{E\left[\left(\frac{c(t+\Delta t)-c(t)}{\Delta t}\right)^2\right]} \text{ and } \sigma_R^2 = \sqrt{E\left[\left(\frac{1}{\Delta t}\log\frac{c(t+\Delta t)}{c(t)}\right)^2\right]} \quad (12)$$

For the NP model, the error distribution for the initial condition of phytoplankton is the same as for the chlorophyll field in the other forward models. For the nutrient field, we assume the mean to be spatially uniform with a value (n=0.3 mmol m^{-3}) prescribed by the domain-average nitrate concentration extracted from the World Ocean Atlas 2005 climatology (Garcia et al., 2006). The covariance for the nutrient initial conditions takes a similar form

$$C_{ij}^{n} = \sigma_n^2 \exp\left(-\frac{|x_i - x_j|}{l_m}\right) \quad (13)$$

and the standard deviation σ_n is assumed to be the same as the mean value (Table 2). The prior distribution for the phytoplankton mortality υ and nutrient uptake rate γ are independent normal distributions with mean 0.1 d^{-1} and 0.3 m^3 mmol^{-1} d^{-1} respectively. These values result in a steady

229 state at the prior mean, consistent with the assumption of zero mean for *S* and *R* in the ADS and
230 ADR models. The prior standard derivation for the uptake and mortality are assumed to be four
231 times their mean value, reflecting large uncertainty in the prior estimates.

232

233 *2.5 Ensemble Kalman Smoother*

234 The EnS algorithm solves the strong constraint data assimilation problem using an
235 analysis scheme and statistical forecasting methodology closely related to the Ensemble Kalman
236 filter (EnKF) described in Evensen (2006). To obtain the model error distributions at the
237 observation points, $f(H\psi_\theta)$, we employ a Monte-Carlo method. For example, in the ADS model
238 spatially variable initial conditions and source-sink terms are simulated from the prior
239 distributions (Equations 10 and 11). The forward model (Equation 2) is integrated with a finite
240 element solver to produce a Monte Carlo sample of the prior model error distribution at the
241 observation points. An analogous procedure is employed for the AD, ADR and NP models.

242 Suppose the model response to the parameters is linear at the observation points,
243 $H\psi_\theta = H\psi_\mu + Q(\theta - \mu)$, where $Q = \frac{\partial H\psi_\theta}{\partial \theta}$. To obtain the optimal estimate of θ we utilize the
244 normal equations,

245 $$0 = \frac{\partial J(\theta)}{\partial \theta} = Q^T W^{-1}(H\psi_\theta - d) + P^{-1}(\theta - \mu) = Q^T W^{-1}(H\psi_\mu + Q(\theta - \mu) - d) + P^{-1}(\theta - \mu) \quad (14)$$

246 Solving for θ we have,

247 $$\theta = \mu + (P^{-1} + Q^T W^{-1} Q)^{-1} Q^T W^{-1}(d - H\psi_\mu) \quad (15)$$

248 Or equivalently, utilizing a matrix lemma,

$$\theta = \mu + PQ^T (QPQ^T + W)^{-1}(d - H\psi_\mu) = \mu + C_{\theta d}(C_{dd} + W)^{-1}(d - H\psi_\mu). \quad (16)$$

Here $C_{\theta d} = E[(\theta - \mu)(H\psi_\theta - H\psi_\mu)]$ and $C_{dd} = E[(H\psi_\theta - H\psi_\mu)(H\psi_\theta - H\psi_\mu)]$ are the model error covariances between the parameters and observation points and the model error covariances between the observation points respectively. The EnS optimal estimate uses a Monte Carlo approximation of these two covariance matrices, thus avoiding the need for a gradient calculation. The optimality of the estimate is conditioned on the existence of a good linear approximation to the dynamic model, though it is never computed explicitly. The approximation only needs to be valid at the observation points in space/time and over the likely regions in the prior distribution for θ. A more detailed derivation of this optimal estimate, its posterior statistics and method for its computation are described in Smith et al. (2009). The posterior estimate of the state is obtained by solving the forward model for a sample of the parameters drawn from their posterior distribution. In this sense, the model provides a stochastically-based strong constraint estimate of the model parameters and state.

2.6 Iterative Ensemble Kalman Smoother

In cases where the log likelihood (Equation 8) is not approximately quadratic we can generalize the EnS approach by iterating the analysis scheme, linearizing the cost function about a series of points of increasing likelihood. The linearization is accomplished with an ensemble approximation to the gradient rather than a numerical or analytic linearization of the forward model. If $H\psi_\theta$ is differentiable, then for any value of the parameter vector *y* we can linearly approximate the cost function in some neighborhood of *y*

$$J(\theta) \cong (H\psi_y + Q_y(\theta - y) - d)^T W^{-1}(H\psi_y + Q_y(\theta - y) - d) + (\theta - \mu)^T P^{-1}(\theta - \mu) \quad (17)$$

271 And thus

272 $$\frac{\partial J(\theta)}{\partial \theta}_{\theta=y} \cong Q_y^T W^{-1}\left(H\psi_y + Q_y(\theta - y) - d\right) + P^{-1}(\theta - \mu) \quad (18)$$

273 where

274 $$Q_y = \frac{\partial H\psi_\theta}{\partial \theta}_{\theta=y} \quad (19)$$

275 is the sensitivity matrix evaluated at $\theta = y$. The first order condition for a minimum is found by

276 setting Equation 17 to zero and solving for θ obtaining,

277 $$\theta = \mu + PQ_y^T\left(Q_y P Q_y^T + W\right)^{-1}\left(d - H\psi_y + Q_y(y - \mu)\right). \quad (20)$$

278 Note that here μ and P are the specified prior mean and covariance for θ rather than their Monte

279 Carlo approximation as in the EnS.

280 We wish to find a sequence of parameter values, $y_1, y_2, ..., y_n$ that will converge to the

281 maximum likelihood estimate for θ. The starting point for this sequence is the prior mean,

282 $y_1 = \mu$. Using the optimal update based on the local normal equations, we define the update

283 candidate

284 $$y'_{i+1} = \mu + PQ_{y_i}^T\left(Q_{y_i} P Q_{y_i}^T + W\right)^{-1}\left(d - H\psi_{y_i} + Q_{y_i}(y_i - \mu)\right) \quad (21)$$

285 Because the linear approximation Q_{y_i} is local and may not be valid out to y'_{i+1}, the update, y_{i+1},

286 is the point on the line between y_i and y'_{i+1} that minimizes the exact cost function (Equation 8).

287 Formally we have $y_{i+1} = (1-\lambda)y_i + \lambda y'_{i+1}$ where

288 $$\lambda = \arg\min(J_i'(\lambda)) \text{ for } J_i'(\lambda) = J((1-\lambda)y_i + \lambda y'_{i+1}). \quad (22)$$

289 The discrete ensemble (of size N_S) over which the minimum of the exact cost function is

computed is $\lambda_j = \frac{j}{N_s}$ for $j = 0,1,..., N_s$. The minimal cost corresponds to the optimal step size. The implementation of the optimal step size calculation utilizes the existing parallel ensemble forward model, though other choices might be more efficient such as a divide and conquer approach or curve fitting.

The local derivative estimates, Q_{y_i}, are computed with an SVD decomposition of an ensemble of parameter vectors, and the solution of the dynamical model for the ensemble. Let

$$\left[\theta_{i,1}, \theta_{i,2}, ..., \theta_{i,n_e}\right] = \left[y_i + \eta_{i,1}, y_i + \eta_{i,2}, y_i + \eta_{i,n_e}\right] = \left[Y_i\right] = \left[U_i\right]\left[D_i\right]\left[V_i\right]^T \quad (23)$$

denote the SVD decomposition of the ensemble of parameter vectors at the i^{th} iteration and let $\left[M_i\right] = H\psi\left(\left[Y_i\right]\right)$ denote the ensemble estimate at the data points. By definition U_i and V_i are unitary and D_i is diagonal. The approximate sensitivity matrix is given by $Q_{y_i} = \left[M_i\right]\left[V_i\right]\left[D_i\right]^{-1}\left[U_i\right]^T$ as in related methods such as the Iterative Ensemble Kalman Filters of Li and Reynolds (2009). The ensemble is regenerated at each iteration; the perturbation vectors $\eta_{i,j}$ are simulated independently from a scaled prior covariance with mean zero. The ensemble standard deviation for the $\eta_{i,j}$ is 1/1000 of the prior standard deviation. The sampling scheme used to generate the ensemble is generally not optimal for estimation of the sensitivity matrix. The problem of defining an optimal sampling strategy for the derivative estimate is left for future work.

The requisite size of the ensemble for estimating Q_{x_i} depends on the number of parameters being estimated. For problems with only a handful of parameters a deterministic approach to sampling, such as the sampling scheme of the Unscented Kalman filter (UKF)

(Julier and Uhlmann, 1997), would be a natural choice. Such a scheme would require a sample size of $1+2\dim(\theta)$. For the joint initial condition and spatially variable parameter estimation problems solved herein, $\dim(\theta) \cong 10000$, making such a sample unfeasible for our numerical experiments.

In addition to dealing with strongly non quadratic log likelihoods, the ItEnS allows the sampling distribution to not conform to the prior distribution. This is advantageous if the prior error distribution is ill specified, such as the assumption of a Gaussian prior for a field which must be positive in the dynamical model. The ItEnS methodology also guarantees convergence to a minimum of the cost function whose basin of attraction contains the prior mean. In cases where the likelihood is multimodal, this may not be a global minimum (e.g. Smith (2007)). However, an extensive search for the global minima can be conducted utilizing multiple starting points. If the prior estimate is reasonable (or equivalently if the observations are noisy and provide little constraint) the algorithm will converge to the global minimum.

2.7 Experimental design

The data set consists of eleven partial images on July 24, August 3,9,12,17,19,21 September 3, 4, 7 and 9 (Figure 2, top row), which are located in an interior subdomain of the regional model (Figure 1). In order to test our data assimilation methodology, we sequentially subdivided this time series of images into nine time windows, each containing three successive images. In each case, the first and last images were assimilated and the middle image was used to evaluate the posterior estimate (Table 1). For the time scales associated with these experiments, the regional domain was large enough that assimilation of data in the interior

subdomain did not involve boundary conditions of the regional model.

Because satellite-based chlorophyll estimates can be contaminated by a variety of atmospheric and oceanic sources, it is difficult to prescribe an appropriate observational error model. We therefore assess the sensitivity of the estimation to the observational error standard deviation by testing ten values of σ_{obs} with a log uniform structure,

$$\sigma_{\text{obs}} = [.05,\ .1,\ .2,\ .4,\ .8,\ 1.6,\ 3.2,\ 6.4,\ 12.8,\ 25.6]\ \text{mg m}^{-3}.$$

3. Results

The observational basis for this study is satellite-based chlorophyll imagery from late July to early September 2006 (Figure 2, top row). Chlorophyll concentrations in late July and early August are generally low overall. In mid-August, enhanced chlorophyll appears in the vicinity of the shelf break (Figure 1), oriented in the northeast to southwest direction; highest concentrations are located in the northeast. By early September, the enhanced chlorophyll disappears, although weak gradients persist along the shelf break.

The best prior estimate (Figure 2, second row) consists of a simulation with the abiotic AD model initialized with the climatological mean chlorophyll concentration for August derived from MODIS data. The climatology contains enhanced chlorophyll in the northwest corner of the domain, and low values elsewhere—and thus bears little resemblance to the observations in July-September 2006. Nevertheless, this forward model simulation without data assimilation constitutes our best prior estimate of the chlorophyll field for all the models: AD, ADS ($S=0$), ADR ($R=0$), and NP (γ, υ chosen so the right hand sides of Equations 4 and 5 are zero).

Data assimilation vastly improves the fit to passive observations for the entire suite of

dynamical models (Figure 2, rows 3-6). The mid-August enhancement of chlorophyll along the shelf break is recovered in each case, albeit to varying degrees (cf. August 19). Also evident are remnants of the high chlorophyll in the northwestern part of the domain present in the best prior estimate, especially during time periods for which observations are lacking in that particular area (e.g. July 24 / August 9, September 4 / September 9).

The inferred biological parameters vary significantly over time, and depend on the underlying model formulation (Figure 3). Buildup of chlorophyll along the shelf break in mid-August is fostered by enhanced growth in that area, reflected by positive $S(x,y)$ and $R(x,y)$ in the ADS and ADR models, respectively (Figure 3, rows 1 and 2). These areas of growth are flanked by areas of mortality (negative $S(x,y)$ and $R(x,y)$), which tend to keep the biomass enhancement confined to the shelf break. Disappearance of the chlorophyll enhancement in late August results from widespread mortality in the ADS and ADR models. Dynamics of the NP model are considerably different (Figure 3, row 3). The mid-August chlorophyll enhancement is bolstered by high nutrients extending seaward from the shelf break. Lower nutrients landward of the shelf break (August 12, 17, and to some extent on August 19) prevent chlorophyll buildup in that area. The decline in biomass along the shelf break from late August to early September is controlled primarily by a decrease in the nutrient uptake rate γ and an increase in mortality υ.

4. Discussion

4.1 Misfit

Fit to the active data depends on both the observational error and the underlying dynamical model (Figure 4). As expected, the fits generally degrade monotonically with

increasing σ_{obs}. However, there are some exceptions (e.g. experiment 6, NP model, observational error 0.1-0.4). These could be due to local minima or premature convergence triggered by the stopping rule (see section 4.3 below) when Monte Carlo errors in the gradient calculation cause an increase in the cost function. For some models (especially the ADR model), there is a systematic tendency for a local maximum in misfit at the lowest observational error. This "convergence error" is likely a result of the Monte Carlo approximation, and could be ameliorated by an increase in ensemble size (with a commensurate impact on computational cost).

On average, the ADR model fits the data better than the ADS model, which fits better than the NP model, which fits better than the AD model. Why are the fits so different amongst the various models? There are three reasons: differences in the number of degrees of freedom, differences in model structure, and differences in the prior distributions of the inferred parameters. For example, the AD model has the fewest number of degrees of freedom, and it produces the worst fit. The ADS and ADR models both have the same number of degrees of freedom, yet the ADR model fits the active data systematically better than the ADS model. Due to the exponential nature of the solution to the ADR model, it is generally more effective at fitting outliers in the terminal data than the linear ADS model. Moreover, the prior distributions of *S* and *R* are necessarily different given they have different units—and those differences undoubtedly affect the fit.

Although the degrees of freedom for the NP model are slightly higher than the ADS and ADR model ($2N_m$+2 rather than $2N_m$), the misfit is generally greater. There are several reasons for this, including the aforementioned differences in specification of prior for *n* relative to *S* and *R*, as well as the positive definite constraint on *n*. Moreover, the nature of the inversion is quite

399 different in the NP model: whereas in the ADS and ADR cases consist of inverting for initial
400 conditions for the single state variable *c* and a spatially variable parameter of the right hand side,
401 in the NP case we invert for initial conditions for the two state variables *n* and *p* plus two
402 parameters that tie them together dynamically. Unlike the inversions for *S* and *R* in the ADS and
403 ADR models, diffusion acts on the inferred initial conditions for *n* in the NP model, leading to
404 fewer effective degrees of freedom in fitting the terminal data. The misfit of the NP model
405 relative to terminal data is further limited by the NP model's tendencies toward a spatially
406 uniform steady state at long times. This last effect becomes more important in the longer
407 simulations (experiments 1, 6 and 7).

408

409 *4.2 Skill*

410 We define the skill of the estimation procedure as the ratio of root mean square (RMS)
411 prior misfit to unassimilated data to the RMS of the posterior misfit to the same data. This
412 metric is non-dimensional and can be compared across the nine time windows which each have
413 different prior misfits to their passive data. If this ratio is greater than one, we consider the
414 estimation procedure to have skill.

415 Averaging the results across all nine time windows, we find that all of the models have
416 skill across the full range of σ_{obs} (Figure 5). Average skill is optimal for σ_{obs}=0.8-1.6 mg m^{-3},
417 depending on the model. Skill is poor across all models and observational error for time
418 windows 1 and 2, and good for all models in for time windows 3, 5, and 6. Overfitting (poor
419 skill at low σ_{obs}) with the ADR model is found in experiments 4 and 9. Overfitting also occurs
420 with the AD model in experiments 7 and 8. The NP model only exhibits overfitting in

421 experiment 8.

422

423 *4.3 Non quadratic log likelihoods: necessity of an iterative approach*

424 To illustrate the necessity of the iterative approach, we evaluate the cost function between
425 the prior estimate ($\theta = \mu$) and the first candidate estimate of the ItEnS ($\theta = y'_2$). The cost
426 function is computed on regularly spaced values in the interval $\mu \le \theta \le y'_2$. We find that the cost
427 function deviations from quadratic vary greatly amongst the nine experiments with each model
428 (Figure 6). As expected, the cost functions for the explicitly nonlinear models (ADR and NP)
429 exhibit the most significant departures from quadratic form. The ADR model exhibits
430 asymmetry about the minimum, while the NP model occasionally contains multiple local
431 minima. In experiment 9 the cost function is quadratic for all models.

432 For most of the experiments we find convergence of the cost function in 1-10 iterations,
433 most requiring only a single iteration due to the cost function being nearly quadratic. Models
434 with strong nonlinearities and low observational error generally required more iteration. We
435 consider the convergence to have occurred if the improvement in the cost function is less than
436 1/1000 of the current value, i.e. $y_{i-1} - y_i < \frac{y_{i-1}}{1000}$.

437

438 **5. Conclusions**

439 We have demonstrated an alternative smoother formulation for strongly non-linear
440 systems, the ItEnS. As in the EnS, the strong constraint data assimilation problem is formulated
441 in a Bayesian framework and solved without the need for a tangent linear model.

442 Bayesian formalism combined with dynamical models provides a useful context for
443 compositing satellite-based ocean color imagery. We find that, with respect to the hindcasting
444 experiments presented here, assimilating chlorophyll data improved the fit to unassimilated data
445 over a broad range of presumed observational error. This is an important property because the
446 relationship between ocean color and phytoplankton abundance is highly variable in both space
447 and time, and consequently error models are rarely specified with great confidence.
448 For the abiotic AD model based on advection and diffusion only, the estimation procedure
449 was used to infer optimal initial conditions. For this model we find an average of 18%
450 improvement in the fit to unassimilated data, demonstrating the utility of assimilating data into
451 circulation-based predictions of surface chlorophyll.
452 We find significant skill in all of the coupled physical-biological models tested here.
453 While the ADR and ADS models generally fit the assimilated data better than the NP model, the
454 skill of the three models was similar. Examination of the results over a range of prescribed
455 observational error σ_{obs} revealed the best improvement in fit to the passive data averaged 36%,
456 43%, and 32% for the ADS, ADR, and NP models respectively. The skill of each biotic model
457 was better than the purely physical advection-diffusion model, and the inferred biological
458 dynamics of course depends on model formulation.
459 Looking deeper than these average statistics, we note that the skill of the assimilation
460 procedure was more dependent on the particular time window being tested than on the
461 underlying dynamical model or presumed observational error. In other words, the results depend
462 strongly on the space-time distribution of the data and their depiction of the oceanographic
463 phenomenology. For example, in some experiments for very low σ_{obs} we find poor skill with the
464 ADR model due to classical overfitting. Thus, although the mean skill scores mentioned above

are promising, the results of individual experiments can be substantially worse. Detailed skill assessment of such methodologies is an essential ingredient to their practical application.

Acknowledgements
We thank Ruoying He for providing the circulation hindcasts used for the inversions. This work was supported by NSF grant DMS-0417845 and ONR grant N00014-06-1-0739.

References

Cullen, J.J., 1982. The deep chlorophyll maximum: comparing vertical profiles of chlorophyll *a*. Can. J. Fish. Aquat. Sci., 39: 791-803.

Evensen, G., 2006. Data Assimilation: the Ensemble Kalman Filter. Springer-Verlag, Berlin Heidelberg: 279 pp.

Fan, W. and Lv, X., 2009. Data assimilation in a simple marine ecosystem model based on spatial biological parameterizations. Ecological Modelling, 220(17): 1997-2008.

Fennel, K., Losch, M., Schroter, J. and Wenzel, M., 2001. Testing a marine ecosystem model: sensitivity analysis and parameter optimization. Journal of Marine Systems, 28: 45-63.

Friedrichs, M.A.M., 2002. Assimilation of JGOFS EqPac and SeaWiFS data into a marine ecosystem model of the central equatorial Pacific Ocean. Deep Sea Research II, 49: 289-319.

Garcia-Gorriz, E., Hoepffner, N. and Ouberdous, M., 2003. Assimilation of SeaWiFS data in a coupled physical-biological model of the Adriatic Sea. Journal of Marine Systems, 40-41: 233-252.

Garcia, H.E., Locarnini, R.A., Boyer, T.P. and Antonov, J.I., 2006. World Ocean Atlas 2005: Nutrients (phosphate, nitrate, silicate), U.S. Government Printing Office, Washington, D.C.

Gregg, W.W., 2008. Assimilation of SeaWiFS ocean chlorophyll data into a three-dimensional global ocean model. Journal of Marine Systems, 69: 205-225.

He, R. and Chen, K., submitted. Investigation of the Northeastern North America Coastal Circulation with a Regional Circulation Hindcast Experiment. Journal of Marine Research.

Hofmann, E.E. and Friedrichs, M.A.M., 2002. Predictive modeling for marine ecosystems. The Sea, 12: 537-565.

Ishizaka, J., 1990. Coupling of Coastal Zone Color Scanner Data to a Physical-Biological Model of the Southeastern United-States Continental-Shelf Ecosystem .3. Nutrient and Phytoplankton Fluxes and Czcs Data Assimilation. Journal of Geophysical Research-Oceans, 95(C11): 20201-20212.

Julier, S. and Uhlmann, K., 1997. A New Extension of the Kalman Filter to Nonlinear Systems, SPIE AeroSense Symposium, Orlando, FL, pp. 182-193.

Li, G. and Reynolds, A.C., 2009. Iterative Ensemble Kalman Filters for Data Assimilation. Society of Petroleum Engineers Journal, 14(3): 496-505.

Lynch, D.R., McGillicuddy Jr, D.J. and Werner, F.E., 2009. Skill assessment for coupled biological/physical models of marine systems. Journal of Marine Systems, 76(1-2): 1-3.

Natvik, L.J. and Evensen, G., 2003a. Assimilation of ocean colour data into a biochemical model of the North Atlantic. Part 1: Data assimilation experiments. Journal of Marine Systems, 40-41: 127-153.

Natvik, L.J. and Evensen, G., 2003b. Assimilation of ocean colour data into a biochemical model of the North Atlantic. Part 2: Statistical Analysis. Journal of Marine Systems, 40-41: 155-169.

Smith, K.W., 2007. Cluster ensemble Kalman filter. Tellus, 59A: 749-757.

Smith, K.W., McGillicuddy Jr, D.J. and Lynch, D.R., 2009. Parameter estimation using an ensemble smoother: The effect of the circulation in biological estimation. Journal of Marine Systems, 76(1-2): 162-170.

van Leeuwen, P.J. and Evensen, G., 1996. Data assimilation and inverse methods in terms of a probabalistic formulation. Monthly Weather Review, 124: 2898-2913.
Zhao, Q. and Lu, X., 2008. Parameter estimation in a three-dimensional marine ecosystem model using the adjoint technique. Journal of Marine Systems, 74(1-2): 443-452.

Figure Captions

Figure 1. Model domain and mean circulation for August 2006 extracted from the He and Chen (submitted) hindcast. Bold line depicts the boundary of the 2°x2° subdomain for the data assimilation experiments. Thin gray lines show the 30, 60, 100, 200, 500, 1000, 2000, 3000 and 4000 meter isobaths.

Figure 2. Top row: Sequence of satellite-based chlorophyll estimates in the 2°x2° subdomain domain bounded by 38-40°N and 72-74°W (indicated by the dashed line in Figure 1). Rows 2-6 depict simulated chlorophyll for various dynamical models at the times for which passive data are available in each of the nine time windows (Table 1). Observational error for this suite of results is σ_{obs} = 0.8 mg m^{-3}, for which skill is at or near maximum in a mean sense (Figure 5, lower right).

Figure 3. Inferred biological parameters for the ADS (top row), ADR (middle row), and NP (bottom row) models. Time series correspond to the results presented in Figure 2. Values of the nutrient uptake (γ) and phytoplankton mortality (υ) parameters inferred for the NP model are reported below each nutrient field (bottom row). Date labels along the top are identical to those in Figure 2, indicating the intermediate dates on which the solution is evaluated with passive data (see text). The inferred initial nutrient concentrations (bottom row) pertain to the start of each experiment, and as such correspond to the dates shown one column to the left. In the case of the leftmost column, the initial nutrient field corresponds to July 24 (Figure 3).

Figure 4. RMS of posterior misfit to active data as a function of observational error for the four dynamical models in each of the nine time windows. The lower right panel is the average of all nine experiments. The dashed line represents RMS of the prior misfit.

Figure 5. Skill in each of the experiments and skill averaged across all nine experiments. Skill is defined as the ratio of RMS prior misfit to unassimilated data to the RMS of the posterior misfit to the same data.

Figure 6. Normalized cost function between prior estimate $(\theta = \mu)$ and first candidate estimate $(\theta = y'_2)$ in the first iteration the ItEnS. The departure of the cost function from quadratic curve depends on the model as well as the data. For a purely quadratic cost function the curves should be quadratic with the minima occurring at the right hand side of the plot. The actual minima are marked as open circles on the curves. The cost functions shown here are for the case $\sigma_{obs} = 0.4$.

Table(s)

Click here to download Table(s): Tables.docx

Tables

Experiment	Active Data	Passive data
1	7/24, 8/9	8/3
2	8/3, 8/12	8/9
3	8/9, 8/17	8/12
4	8/12, 8/19	8/17
5	8/17, 8/21	8/19
6	8/19, 9/3	8/21
7	8/21, 9/4	9/3
8	9/3, 9/7	9/4
9	9/4, 9/9	9/7

Table 1: Dates of images used in the nine experiments used to test the assimilation procedure.

Parameter	Value
Observational error length scale, l_{obs}	10 km
Model error length scale, l_m	34 km
Chlorophyll/Phytoplankton scaled standard error, σ_0	1.6
Source-Sink standard error, σ_s	0.5 m^3 $mmol^{-1}$ d^{-1}
Growth rate standard error, σ_R	0.5 d^{-1}
Nutrient scaled standard error, σ_n	0.3 mmol m^{-3}

Table 2: Hyper-parameters for the prior models and values used in the assimilation experiments.

Figure 1

Click here to download Figure(s): figure1_domainv3.eps

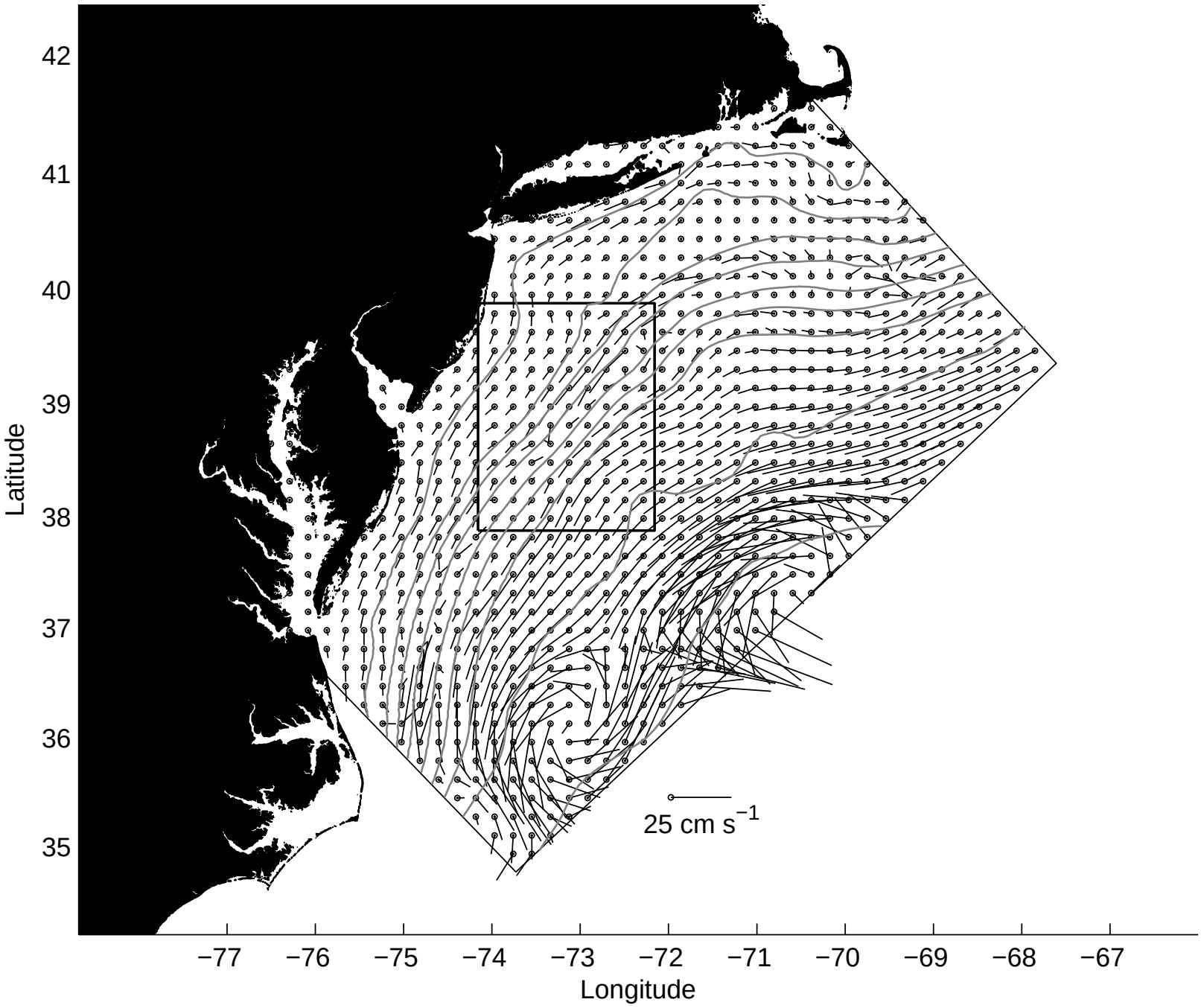

Figure 2

Click here to download Figure(s): figure2_prediction.s8.eps

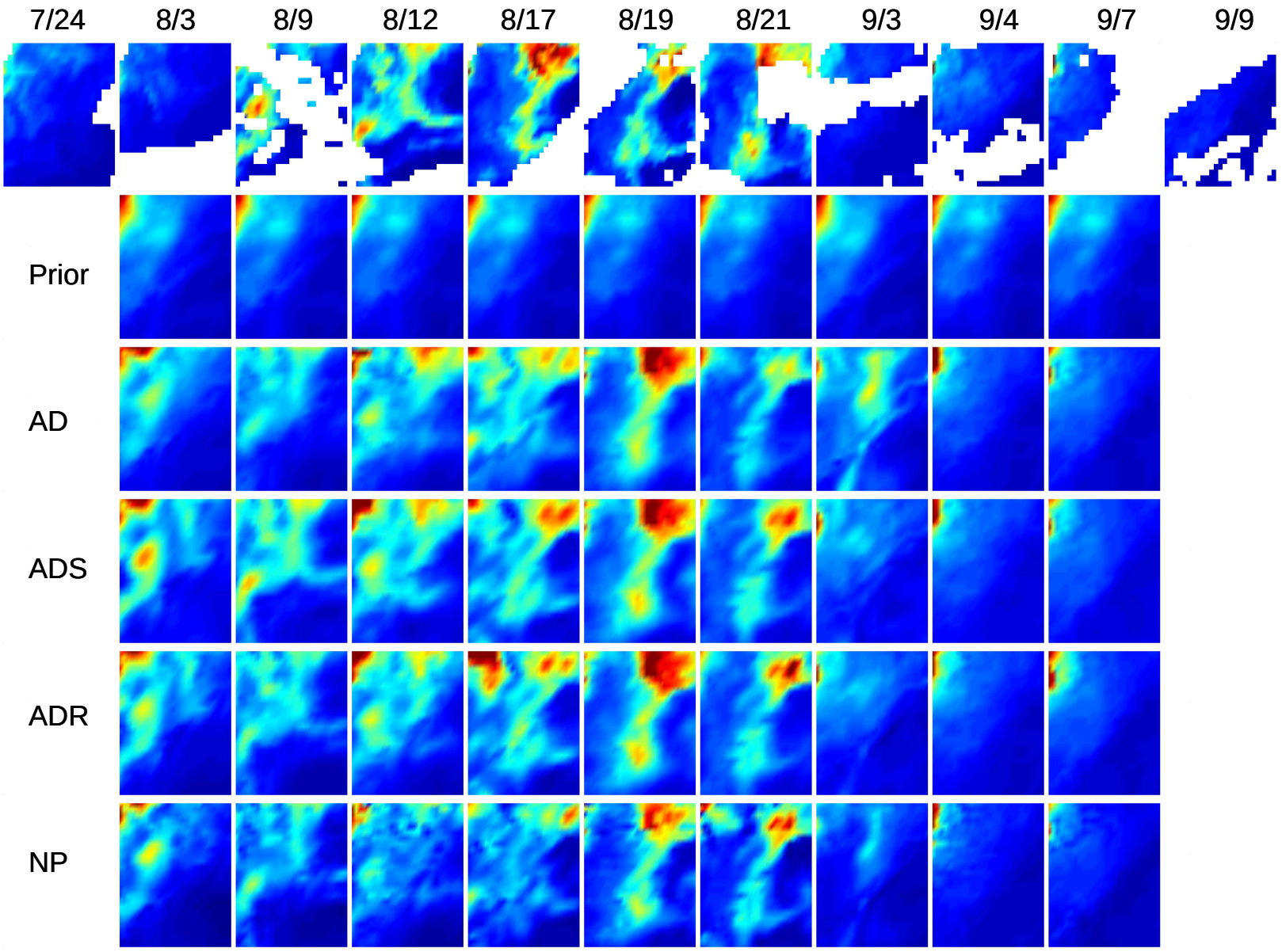

Figure 3

Click here to download Figure(s): figure3_BioICPrediction.s8_v2.eps

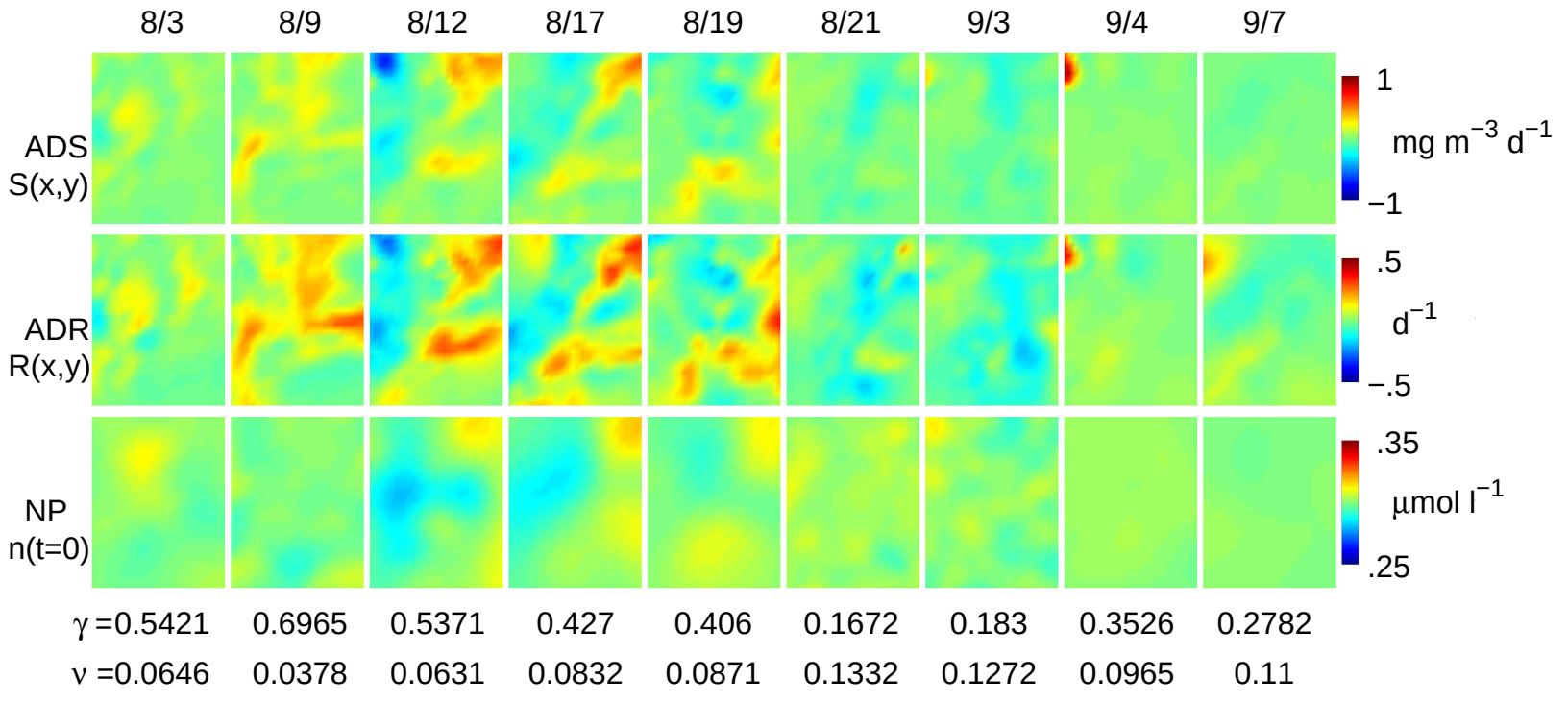

Figure 4

Click here to download Figure(s): figure4_misfits.eps

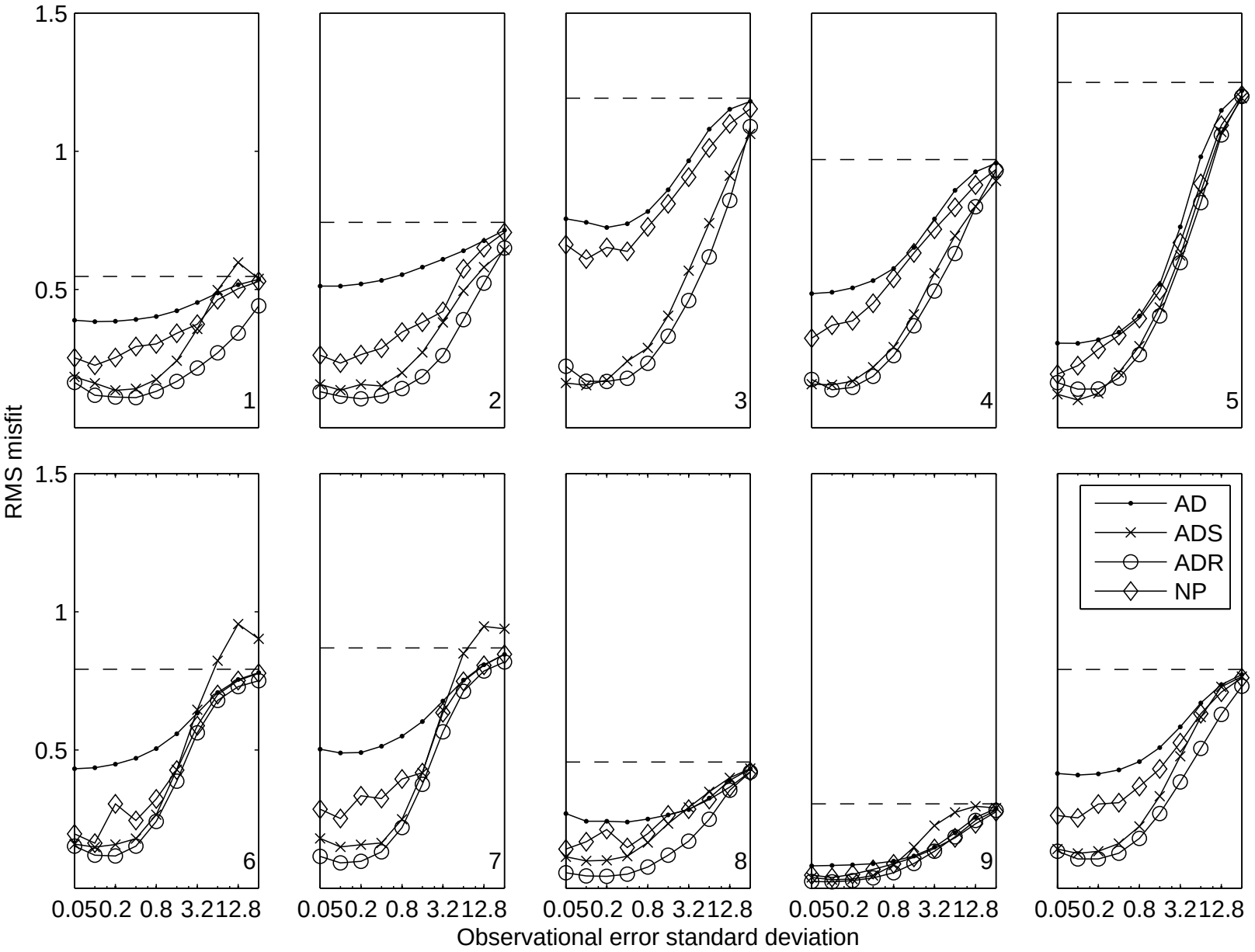

Figure 5

Click here to download Figure(s): figure5_skill.eps

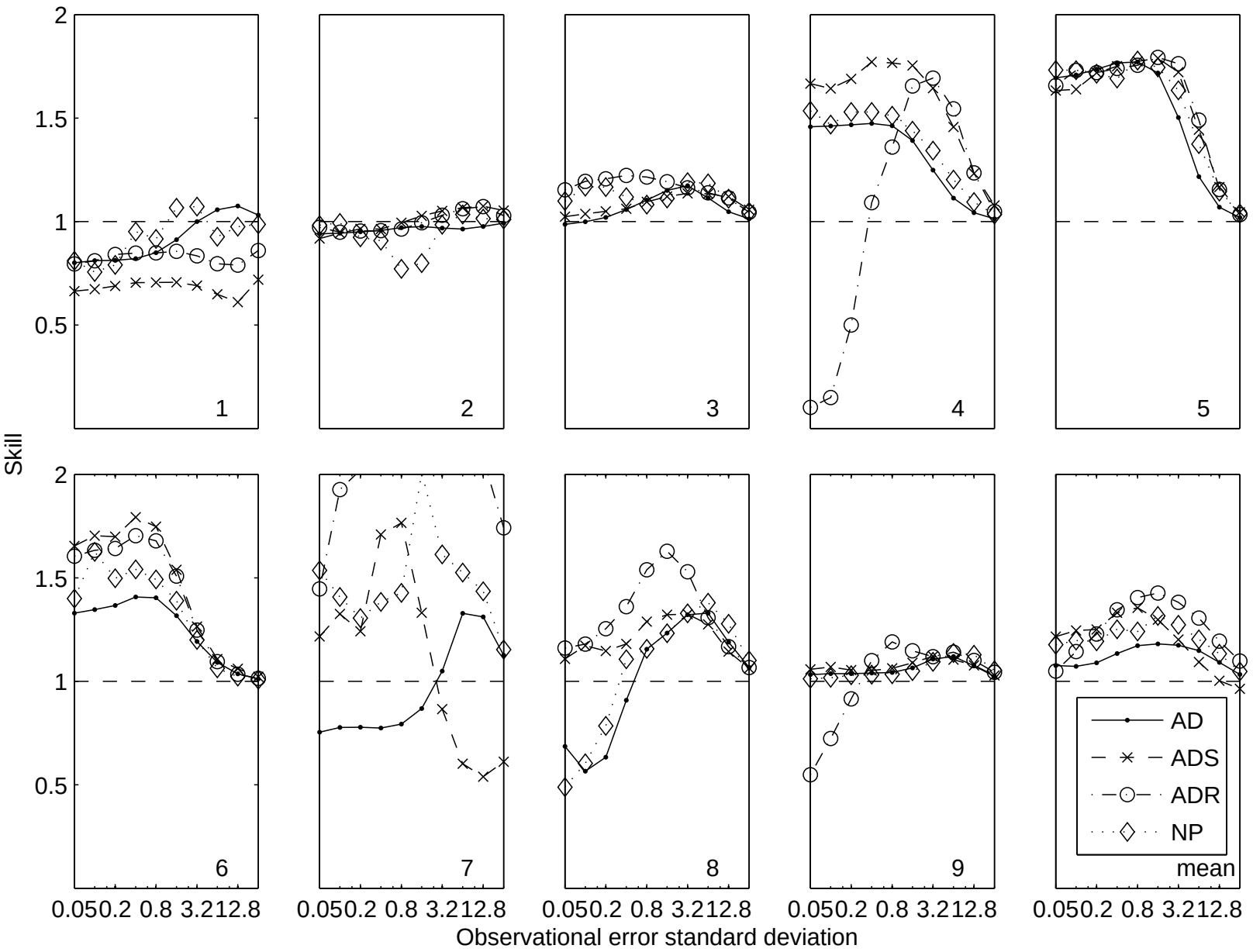

Figure 6

Click here to download Figure(s): figure6_nlcost.eps